

# Objective Bayes Covariate-Adjusted Sparse Graphical Model Selection<sup>§</sup>

Guido Consonni

Università Cattolica del Sacro Cuore

`guido.consonni@unicatt.it`

Luca La Rocca ¶

Università di Modena e Reggio Emilia

`luca.larocca@unimore.it`

October 9, 2015

<sup>§</sup>*Running headline:* Covariate-Adjusted Sparse DAG Selection

¶Corresponding author

## Abstract

We present an objective Bayes method for covariance selection in Gaussian multivariate regression models whose error term has a covariance structure which is Markov with respect to a Directed Acyclic Graph (DAG). The scope is covariate-adjusted sparse graphical model selection, a topic of growing importance especially in the area of genetical genomics (eQTL analysis). Specifically, we provide a closed-form expression for the marginal likelihood of any DAG (with small parent sets) whose computation virtually requires no subjective elicitation by the user and involves only conjugate matrix normal Wishart distributions. This is made possible by a specific form of prior assignment, whereby only one prior under the complete DAG model need be specified, based on the notion of fractional Bayes factor. All priors under the other DAG models are derived using prior modularity, and global parameter independence, in the terminology of Geiger & Heckerman (2002). Since the marginal likelihood we obtain is constant within each class of Markov equivalent DAGs, our method naturally specializes to covariate-adjusted decomposable graphical models.

*Keywords:* Bayesian model selection; Covariate-adjusted graphical model; Covariance selection; Decomposable graphical model; Directed acyclic graphical model; Fractional Bayes factor; Gaussian graphical model; Gaussian multivariate regression; Marginal likelihood; Sparse model selection.

# 1 Introduction

Graphical models are a well-established tool in multivariate statistics. They allow to simplify high-dimensional distributions, both in terms of computations and in terms of interpretation, on the basis of a graph representing independencies between variables. We assume the reader is familiar with the basic theory of undirected and acyclic directed graphical models, as presented for instance in Cowell *et al.* (1999), or Lauritzen (1996); see also Whittaker (1990).

Our interest lies in a collection of  $q$  random variables whose joint distribution, having density with respect to a product measure, embodies a conditional independence structure which can be represented by a Directed Acyclic Graph (DAG). This means that each variable is conditionally independent of its non-descendants given its parents; see Cowell *et al.* (1999, sect. 5.3). Such a distribution is said to be Markov with respect to the DAG. A DAG model is a (parametric) family of multivariate distributions which are Markov with respect to a DAG. We will consider in particular Gaussian DAG models. Then, the DAG structure will be reflected in the covariance matrix  $\Sigma$ : if the DAG is complete,  $\Sigma$  will be unconstrained; for an incomplete DAG,  $\Sigma$  will present constrained entries. Notice that an unconstrained covariance matrix still has to be s.p.d. (symmetric positive definite).

Typically, the DAG structure is unknown, and we want to infer it from  $n$  joint observations of the  $q$  variables. From a Bayesian viewpoint one starts with a prior distribution on the collection of all DAGs (prior on model space), as well as with a prior distribution on the parameter space of each given DAG (parameter prior). Given these prior inputs, Bayesian inference produces a posterior probability on the space of all DAGs, which summarizes all the uncertainty in the light of the available data. Several papers have addressed this problem for the case in which the  $n$  observations are i.i.d. (independent and identically distributed) conditionally on the parameters of the model; see for instance Dawid & Lauritzen (1993); Spiegelhalter *et al.* (1993); Heckerman *et al.* (1995); Madigan *et al.* (1996). Of special interest for this paper is the work by Geiger & Heckerman (2002); see also Consonni & La Rocca (2012)

and Kuipers *et al.* (2014) for corrections. Geiger & Heckerman (2002) listed a set of assumptions on the collection of parameter priors (across DAGs) which permit their construction starting from a single parameter prior under a complete DAG (a DAG with all pairs of vertices directly connected). This represents a dramatic simplification because: i) the specification of only one distribution is required, while all the remaining priors are derived from this one; ii) the latter distribution is placed on an unconstrained parameter space describing the model with no independencies. In the Gaussian case ii) means that one can use a standard Inverse Wishart on the covariance matrix, equivalently a Wishart on the corresponding precision matrix (defined as the inverse of the covariance matrix) so that the marginal likelihood can be expressed in closed form.

Different DAGs may define the same DAG model, in which case they are called Markov equivalent. Accordingly, the set of all DAGs for the  $q$  variables can be partitioned into Markov equivalence classes (corresponding to distinct DAG models). If DAGs are meant to specify exclusively conditional independencies, as opposed to causal relationships (Lauritzen, 2001; Dawid, 2003), then all DAGs specifying the same set of conditional independencies should be regarded as indistinguishable using observational data. The method by Geiger & Heckerman (2002) ensures that DAGs belonging to the same equivalence class obtain the same marginal likelihood. As a consequence, their method can also be used to infer decomposable graph structures, by simply replacing each structure with an equivalent DAG (no matter which).

Despite its many advantages, the inferential procedure proposed by Geiger & Heckerman (2002) still requires the specification of a potentially high-dimensional parameter prior (especially in large  $q$  settings). This naturally suggests an objective Bayes approach, which is virtually free from prior elicitation. We carried out this program in Consonni & La Rocca (2012) for Gaussian DAG models, using the method of the fractional Bayes factor (O’Hagan, 1995). Our findings were consistent with, and extended, those presented in Carvalho & Scott (2009) for Gaussian decomposable graphical models, which relied on the use of the hyper-inverse Wishart distribution (Letac & Massam, 2007).

More recently, research has shifted towards *covariate-adjusted* estimation of covariance matrices. Motivation for this research stems from the analysis of genetical genomics data (eQTL analysis) where the aim is to study conditional dependence structures of gene expressions after the confounding genetic effects are taken into account. Indeed, an important finding from many genetical genomics experiments is that the gene expression level of many genes is inheritable and can be partially explained by genetic variation; see e.g. Brem & Kruglyak (2005). Since some genetic variants have effects on the expression of multiple genes, they act as confounders when trying to learn the association between the genes. Accordingly, ignoring the effects of genetic variants on the gene expression levels can lead to both false positive and false negative associations in the gene network graph. The effect of genetic variants on gene expression therefore needs to be adjusted in estimating the high-dimensional precision matrix (Cai *et al.*, 2013; Chen *et al.*, 2013).

The problem is usually formulated as one of joint estimation of multiple regression coefficients and a precision matrix, with the latter assumed to be Markov with respect to some graph. Since these models are used in high-dimensional settings, both the regression and the covariance structure are assumed to be sparse. Rothman *et al.* (2010), Yin & Li (2011) and Chen *et al.* (2013) assume that the error term is multivariate normal; this assumption is relaxed in the paper by Cai *et al.* (2013). The literature in the area, as exemplified in all the papers above, is carried out within a constrained minimization approach (under a suitable norm). Contributions in the Bayesian framework are still very limited. A notable exception is Bhadra & Mallick (2013) who perform variable and covariance selection jointly, using decomposable graphs and weakly informative hierarchical priors.

In this paper we deal with covariate-adjusted selection of Gaussian DAG models within an objective Bayes framework. Specifically, we reconsider the foundations of the approach by Geiger & Heckerman (2002), originally presented for the case of i.i.d. sampling, and show that it can be meaningfully extended to the multivariate regression setting. We provide closed-form expressions for the marginal likelihood of any DAG, then we propose an objective Bayes procedure, based on the fractional

Bayes factor, which works for DAGs with small parent sets. Our results extend to the regression setup those of Consonni & La Rocca (2012) and Carvalho & Scott (2009); they also complement those of Bhadra & Mallick (2013), both because they are derived within an objective framework, and because they cope with a broader family of graphs.

The paper is organized as follows. Section 2 reviews the matrix distributions used in the paper, and section 3 presents the Gaussian multivariate regression setup. Section 4 illustrates our objective framework, while section 5 contains our proposal for covariance selection. Finally, section 6 briefly discusses our work.

## 2 Matrix distributions

Consider  $n$  independent observations on  $q$  continuous dependent variables, arranged in an  $n \times q$  response matrix:

$$\mathbf{Y} = \begin{pmatrix} \mathbf{y}_1^\top \\ \vdots \\ \mathbf{y}_n^\top \end{pmatrix} = \begin{pmatrix} \mathbf{Y}_1 & \dots & \mathbf{Y}_q \end{pmatrix}, \quad (1)$$

where  $\mathbf{y}_i = (y_{i1}, \dots, y_{iq})^\top$  is the  $i$ -th observation, and  $\mathbf{Y}_j = (y_{1j}, \dots, y_{nj})^\top$  represents the observations on the  $j$ -th variable. Let  $\mathbf{X}$  be a design matrix with  $n$  rows and  $p+1$  columns ( $p$  predictors plus intercept) which we assume known without error; denote by  $\mathbf{x}_1^\top, \dots, \mathbf{x}_n^\top$  its rows. We model the observations as  $\mathbf{y}_i | \mathbf{B}, \mathbf{\Sigma} \sim \mathcal{N}_q(\mathbf{B}^\top \mathbf{x}_i, \mathbf{\Sigma})$ , independently over  $i = 1, \dots, n$ , where  $\mathbf{B}$  is an unconstrained  $(p+1) \times q$  matrix,  $\mathbf{\Sigma}$  is an s.p.d.  $q \times q$  matrix, and  $\mathcal{N}_q(\boldsymbol{\mu}, \mathbf{\Sigma})$  denotes the  $q$ -variate normal distribution with mean vector  $\boldsymbol{\mu}$  and covariance matrix  $\mathbf{\Sigma}$ . The  $j$ -th column of  $\mathbf{B}$ , namely  $\mathbf{B}_j$ , is the vector of regression coefficients for the  $j$ -th variable, and  $\mathbb{E}(\mathbf{Y} | \mathbf{B}, \mathbf{\Sigma}) = \mathbf{XB}$ . The distribution of  $\mathbf{Y}$ , given  $\mathbf{B}$  and  $\mathbf{\Sigma}$ , is a special case of the matrix normal distribution; the general case, reviewed in section 2.1, will give a conjugate prior for  $\mathbf{B}$  (given  $\mathbf{\Sigma}$ ). A conjugate prior for  $\mathbf{\Sigma}^{-1}$  will be given by the Wishart distribution, which is reviewed in section 2.2.

## 2.1 Matrix normal

We say that the random matrix  $\mathbf{Y}$  follows the *matrix normal distribution* with mean matrix  $\mathbf{M}$ , row covariance matrix  $\mathbf{\Phi}$ , and column covariance matrix  $\mathbf{\Sigma}$ , when  $\text{vec}(\mathbf{Y})$  follows the multivariate normal distribution with mean vector  $\text{vec}(\mathbf{M})$  and covariance matrix  $\mathbf{\Sigma} \otimes \mathbf{\Phi}$ ; recall that  $\text{vec}(\mathbf{Y})$  is the vector obtained from  $\mathbf{Y}$  by stacking its columns on top of one another, while  $\otimes$  denotes the Kronecker product. If  $\mathbf{Y}$  is an  $n \times q$  matrix,  $\mathbf{M}$  will be an  $n \times q$  matrix,  $\mathbf{\Phi}$  an s.p.d.  $n \times n$  matrix,  $\mathbf{\Sigma}$  an s.p.d.  $q \times q$  matrix, and we will write

$$\mathbf{Y} \mid \mathbf{M}, \mathbf{\Phi}, \mathbf{\Sigma} \sim \mathcal{N}_{n,q}(\mathbf{M}, \mathbf{\Phi}, \mathbf{\Sigma}); \quad (2)$$

see Gupta & Nagar (2000, p. 55), and Dawid (1981), for more information. We obtain the special case where  $\mathbf{Y}$  is a response matrix, previously described and taken up in section 3, by letting  $\mathbf{M} = \mathbf{XB}$ , and  $\mathbf{\Phi} = \mathbf{I}_n$ , where  $\mathbf{I}_n$  is the  $n \times n$  identity matrix.

Let  $\phi_{ii'}$  denote the generic element of  $\mathbf{\Phi}$ , and  $\sigma_{jj'}$  the generic element of  $\mathbf{\Sigma}$ . Clearly, we have  $\mathbb{E}(\mathbf{Y} \mid \mathbf{M}, \mathbf{\Phi}, \mathbf{\Sigma}) = \mathbf{M}$ . Moreover, we have  $\text{Cov}(y_{ij}, y_{i'j'} \mid \mathbf{M}, \mathbf{\Phi}, \mathbf{\Sigma}) = \phi_{ii'} \sigma_{jj'}$ , so that  $\text{Var}(\mathbf{y}_i \mid \mathbf{M}, \mathbf{\Phi}, \mathbf{\Sigma}) = \phi_{ii} \mathbf{\Sigma}$ ,  $i = 1, \dots, n$ , whereas  $\text{Var}(\mathbf{Y}_j \mid \mathbf{M}, \mathbf{\Phi}, \mathbf{\Sigma}) = \sigma_{jj} \mathbf{\Phi}$ ,  $j = 1, \dots, q$ , with  $\text{Var}(\mathbf{u})$  denoting the covariance matrix of the random vector  $\mathbf{u}$ . More generally, we find  $\text{Cov}(\mathbf{y}_i, \mathbf{y}_{i'} \mid \mathbf{M}, \mathbf{\Phi}, \mathbf{\Sigma}) = \phi_{ii'} \mathbf{\Sigma}$  and  $\text{Cov}(\mathbf{Y}_j, \mathbf{Y}_{j'} \mid \mathbf{M}, \mathbf{\Phi}, \mathbf{\Sigma}) = \sigma_{jj'} \mathbf{\Phi}$ , if we denote by  $\text{Cov}(\mathbf{u}, \mathbf{v})$  the cross-covariance matrix of  $\mathbf{u}$  and  $\mathbf{v}$ , whose elements are the covariances between all pairs consisting of one element in  $\mathbf{u}$  and the other in  $\mathbf{v}$ . Notice that  $\text{Cov}(\mathbf{u}, \mathbf{u}) = \text{Var}(\mathbf{u})$ .

Reparameterizing from  $\mathbf{\Sigma}$  s.p.d. to  $\mathbf{\Omega} = \mathbf{\Sigma}^{-1}$  s.p.d., and from  $\mathbf{\Phi}$  s.p.d. to  $\mathbf{K} = \mathbf{\Phi}^{-1}$  s.p.d., which we will find useful for Bayesian analysis, the density of the matrix normal distribution  $\mathcal{N}_{n,q}(\mathbf{M}, \mathbf{K}^{-1}, \mathbf{\Omega}^{-1})$  can be written as

$$f(\mathbf{Y} \mid \mathbf{M}, \mathbf{K}, \mathbf{\Omega}) = \frac{|\mathbf{K}|^{\frac{q}{2}} |\mathbf{\Omega}|^{\frac{n}{2}}}{(2\pi)^{\frac{nq}{2}}} \exp \left\{ -\frac{1}{2} \text{tr}(\mathbf{\Omega}(\mathbf{Y} - \mathbf{M})^\top \mathbf{K}(\mathbf{Y} - \mathbf{M})) \right\}, \quad (3)$$

where  $|\mathbf{\Psi}|$  denotes the determinant of the matrix  $\mathbf{\Psi}$ , and  $\text{tr}(\mathbf{\Psi})$  its trace. Formula (3) follows from the density of  $\text{vec}(\mathbf{Y}) \mid \text{vec}(\mathbf{M}), \mathbf{\Omega}^{-1} \otimes \mathbf{K}^{-1}$ , keeping into account that  $\text{tr}(\mathbf{\Omega} \mathbf{\Psi} \mathbf{K} \mathbf{\Psi}^\top) = \text{tr}(\mathbf{\Psi}^\top \mathbf{\Omega} \mathbf{\Psi} \mathbf{K})$  is the value at  $(\mathbf{\Psi}, \mathbf{\Psi})$  of the bilinear form associated to  $\mathbf{\Omega} \otimes \mathbf{K} = (\mathbf{\Omega}^{-1} \otimes \mathbf{K}^{-1})^{-1}$ , which is the precision matrix of  $\text{vec}(\mathbf{Y})$ , and that  $|\mathbf{\Omega} \otimes \mathbf{K}| = |\mathbf{\Omega}|^n |\mathbf{K}|^q$ ; see Lauritzen (1996, App. B). We call  $\mathbf{K}$  the row precision

matrix of  $\mathbf{Y}$ , and  $\mathbf{\Omega}$  its column precision matrix. Clearly, whenever  $\mathbf{Y} \mid \mathbf{M}, \mathbf{K}, \mathbf{\Omega} \sim N_{n,q}(\mathbf{M}, \mathbf{K}^{-1}, \mathbf{\Omega}^{-1})$ , we have  $\mathbf{Y}^\top \mid \mathbf{M}, \mathbf{K}, \mathbf{\Omega} \sim \mathcal{N}_{q,n}(\mathbf{M}^\top, \mathbf{\Omega}^{-1}, \mathbf{K}^{-1})$ , which means  $\text{vec}(\mathbf{Y}^\top) \mid \mathbf{M}, \mathbf{K}, \mathbf{\Omega} \sim \mathcal{N}_{qn}(\text{vec}(\mathbf{M}^\top), \mathbf{K}^{-1} \otimes \mathbf{\Omega}^{-1})$ .

Now let  $J$  be a proper subset of  $\{1, \dots, q\}$ , and denote by  $\mathbf{Y}_J$  the submatrix of  $\mathbf{Y}$  consisting of the columns indexed by  $J$ . It is immediate to check that  $\text{vec}(\mathbf{Y}_J)$  is multivariate normal with mean vector  $\text{vec}(\mathbf{M}_J)$  and covariance matrix  $\mathbf{\Sigma}_{JJ} \otimes \mathbf{\Phi}$ , where  $\mathbf{\Sigma}_{JJ}$  is the submatrix of  $\mathbf{\Sigma}$  consisting of the rows and columns indexed by  $J$ ; see Lauritzen (1996, prop. (C.4)). Hence, *column marginalization* results in

$$\mathbf{Y}_J \mid \mathbf{M}, \mathbf{\Phi}, \mathbf{\Sigma} \sim \mathcal{N}_{n,|J|}(\mathbf{M}_J, \mathbf{\Phi}, \mathbf{\Sigma}_{JJ}). \quad (4)$$

Notice that, if  $\mathbf{M} = \mathbf{X}\mathbf{B}$ , then  $\mathbf{M}_J = \mathbf{X}\mathbf{B}_J$ .

Finally, letting  $\bar{J} = \{1, \dots, q\} \setminus J$ , it is well known that  $\text{vec}(\mathbf{Y}_J) \mid \text{vec}(\mathbf{Y}_{\bar{J}})$  is multivariate normal with mean vector  $\text{vec}(\mathbf{M}_J) - (\mathbf{\Omega}_{JJ}^{-1} \otimes \mathbf{K}^{-1})(\mathbf{\Omega}_{J\bar{J}} \otimes \mathbf{K})\text{vec}(\mathbf{Y}_{\bar{J}} - \mathbf{M}_{\bar{J}})$ , and precision matrix  $\mathbf{\Omega}_{JJ} \otimes \mathbf{K}$ , where  $\mathbf{\Omega}_{JJ}^{-1} = (\mathbf{\Omega}_{JJ})^{-1}$ ; see Lauritzen (1996, prop. C.5). Since  $(\mathbf{\Omega}_{JJ}^{-1} \otimes \mathbf{K}^{-1})(\mathbf{\Omega}_{J\bar{J}} \otimes \mathbf{K}) = (\mathbf{\Omega}_{JJ}^{-1} \mathbf{\Omega}_{J\bar{J}}) \otimes (\mathbf{K}^{-1} \mathbf{K}) = (\mathbf{\Omega}_{JJ}^{-1} \mathbf{\Omega}_{J\bar{J}}) \otimes \mathbf{I}_n$ , we find

$$\mathbf{Y}_J \mid \mathbf{Y}_{\bar{J}}, \mathbf{M}, \mathbf{K}, \mathbf{\Omega} \sim \mathcal{N}_{n,|J|}(\mathbf{M}_J - (\mathbf{Y}_{\bar{J}} - \mathbf{M}_{\bar{J}})\mathbf{\Omega}_{\bar{J}\bar{J}}\mathbf{\Omega}_{JJ}^{-1}, \mathbf{K}^{-1}, \mathbf{\Omega}_{JJ}^{-1}) \quad (5)$$

for *column conditioning*. In the case  $\mathbf{K} = \mathbf{I}_n$ , formula (5) returns  $\mathbf{y}_{iJ} \mid \mathbf{M}, \mathbf{K}, \mathbf{\Omega} \sim \mathcal{N}_{|J|}(\mathbf{m}_{iJ} - \mathbf{\Omega}_{JJ}^{-1} \mathbf{\Omega}_{J\bar{J}}(\mathbf{y}_{i\bar{J}} - \mathbf{m}_{i\bar{J}}), \mathbf{\Omega}_{JJ}^{-1})$ , independently over  $i = 1, \dots, n$ , where  $\mathbf{y}_{iJ}$  and  $\mathbf{m}_{iJ}$  are the subvectors of  $\mathbf{y}_i$  and  $\mathbf{m}_i$ , respectively, consisting of the elements indexed by  $J$ , while  $\mathbf{m}_i^\top$  is the  $i$ -th row of  $\mathbf{M}$ .

## 2.2 Wishart

Let  $\mathbf{\Omega}$  be a  $q \times q$  *unconstrained* s.p.d. random matrix. We will write  $\mathbf{\Omega} \sim \mathcal{W}_q(a, \mathbf{R})$  to mean that  $\mathbf{\Omega}$  follows a Wishart distribution with density

$$p(\mathbf{\Omega}) = \frac{1}{2^{\frac{aq}{2}} \Gamma_q(\frac{a}{2})} |\mathbf{R}|^{\frac{a}{2}} |\mathbf{\Omega}|^{\frac{a-q-1}{2}} \exp \left\{ -\frac{1}{2} \text{tr}(\mathbf{\Omega}\mathbf{R}) \right\}, \quad (6)$$

$\mathbf{\Omega}$  s.p.d., and  $p(\mathbf{\Omega}) = 0$ , otherwise, where  $\mathbf{R}$  is a  $q \times q$  s.p.d. matrix,  $a$  is a scalar strictly greater than  $q - 1$ , and  $\Gamma_q(\frac{a}{2}) = \pi^{\frac{q(q-1)}{4}} \prod_{j=1}^q \Gamma(\frac{a}{2} + \frac{1-j}{2})$  is the  $q$ -dimensional gamma function at  $a/2$  (generalizing  $\Gamma(a/2) = \int_0^\infty z^{\frac{a}{2}-1} e^{-z} dz$ ). As for parameter



interpretation, it can be shown that  $\mathbb{E}[\boldsymbol{\Omega}|\mathbf{R}, a] = a\mathbf{R}^{-1}$ . Our notation  $\mathcal{W}_q(a, \mathbf{R})$  for the density (6) is essentially that of DeGroot (1970, p. 59); other authors (Press, 1982; Lauritzen, 1996) would use  $\mathbf{R}^{-1}$  in place of  $\mathbf{R}$ .

We now recall some useful results. Let  $\boldsymbol{\Omega}$  be the precision matrix of  $\mathbf{y} | \boldsymbol{\mu}, \boldsymbol{\Sigma} \sim \mathcal{N}_q(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ , that is,  $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$ . Think of  $\mathbf{y}$  as the generic row of the matrix  $\mathbf{Y}$  (dropping subscript  $i$ ). Partition  $\boldsymbol{\Sigma}$  and  $\boldsymbol{\Omega}$  into the blocks corresponding to the variables indexed by  $J$  and its complement  $\bar{J}$ , for a given proper subset  $J$  of  $\{1, \dots, q\}$ :

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{JJ} & \boldsymbol{\Sigma}_{J\bar{J}} \\ \boldsymbol{\Sigma}_{\bar{J}J} & \boldsymbol{\Sigma}_{\bar{J}\bar{J}} \end{bmatrix}, \quad \boldsymbol{\Omega} = \begin{bmatrix} \boldsymbol{\Omega}_{JJ} & \boldsymbol{\Omega}_{J\bar{J}} \\ \boldsymbol{\Omega}_{\bar{J}J} & \boldsymbol{\Omega}_{\bar{J}\bar{J}} \end{bmatrix}. \quad (7)$$

The block  $\boldsymbol{\Sigma}_{JJ}$  is the marginal covariance matrix of  $\mathbf{y}_J$  (obtained from  $\mathbf{y}$  by selecting the elements of  $\mathbf{y}$  indexed by  $J$ ). Denote by  $\boldsymbol{\Sigma}_{JJ.\bar{J}}$  the conditional covariance matrix  $\text{Var}(\mathbf{y}_J | \mathbf{y}_{\bar{J}})$  of  $\mathbf{y}_J$  given  $\mathbf{y}_{\bar{J}}$  (obtained from  $\mathbf{y}$  by complementary selection). Then

$$\boldsymbol{\Sigma}_{JJ.\bar{J}} = \boldsymbol{\Sigma}_{JJ} - \boldsymbol{\Sigma}_{J\bar{J}}\boldsymbol{\Sigma}_{\bar{J}\bar{J}}^{-1}\boldsymbol{\Sigma}_{\bar{J}J} = \boldsymbol{\Omega}_{JJ}^{-1}, \quad (8)$$

that is,  $\boldsymbol{\Sigma}_{JJ.\bar{J}}$  is the *Schur complement* of  $\boldsymbol{\Sigma}_{\bar{J}\bar{J}}$  in  $\boldsymbol{\Sigma}$ , as well as the inverse of  $\boldsymbol{\Omega}_{JJ}$ .

Formula (8) expresses a relationship between four blocks of  $\boldsymbol{\Sigma}$  and a corresponding block of  $\boldsymbol{\Sigma}^{-1} = \boldsymbol{\Omega}$ . Hence, by switching the roles of  $\boldsymbol{\Sigma}$  and  $\boldsymbol{\Omega}$ , we obtain

$$\boldsymbol{\Sigma}_{JJ} = (\boldsymbol{\Omega}_{JJ} - \boldsymbol{\Omega}_{J\bar{J}}\boldsymbol{\Omega}_{\bar{J}\bar{J}}^{-1}\boldsymbol{\Omega}_{\bar{J}J})^{-1} = \boldsymbol{\Omega}_{JJ.\bar{J}}^{-1}, \quad (9)$$

where  $\boldsymbol{\Omega}_{JJ.\bar{J}}^{-1}$  is to be interpreted as Schur complementation followed by inversion. Therefore, working with covariance matrices, marginalization corresponds to submatrix extraction and conditioning to Schur complementation, whereas, working with precision matrices, marginalization corresponds to Schur complementation and conditioning to submatrix extraction.

Now let  $\boldsymbol{\Omega} \sim \mathcal{W}_q(a, \mathbf{R})$ , with  $\mathbf{R}$  an s.p.d. matrix and  $a > q - 1$ . If  $\boldsymbol{\Omega}$  is partitioned as in (7), and  $\mathbf{R}$  is partitioned accordingly, then

$$\boldsymbol{\Omega}_{JJ.\bar{J}} \sim \mathcal{W}_{|J|}(a - |\bar{J}|, \mathbf{R}_{JJ}), \quad (10)$$

independently of  $(\boldsymbol{\Omega}_{J\bar{J}}, \boldsymbol{\Omega}_{\bar{J}\bar{J}})$ , where of course  $|\bar{J}| = q - |J|$ ; see Lauritzen (1996, prop. C.15) who also gives the distribution of  $(\boldsymbol{\Omega}_{J\bar{J}}, \boldsymbol{\Omega}_{\bar{J}\bar{J}})$ .

### 3 Gaussian multivariate regression

We return to the scenario discussed in the Introduction, leading to covariate-adjusted graphical model selection, and to the response matrix  $\mathbf{Y}$  introduced at the beginning of section 2. Denote by  $\mathbf{Z}$  the  $n \times p_\star$  matrix of all possible  $p_\star$  predictors. In eQTL analysis  $p_\star$  is typically very large, and often much larger than  $n$ . However, because of sparsity considerations, only models of the type  $\mathbf{Y} = \mathbf{X}\mathbf{B} + \mathbf{E}$  need be taken into consideration, where  $\mathbf{X}$  is an  $n \times (p + 1)$  design matrix having the unit vector  $\mathbf{1}_n$  as first column and  $p \ll p_\star$  predictors from  $\mathbf{Z}$  as remaining columns, while  $\mathbf{E}$  is an  $n \times q$  matrix of error terms with distribution  $\mathcal{N}_{n,q}(\mathbf{0}, \mathbf{I}_n, \mathbf{\Omega}^{-1})$ , and  $\mathbf{B}$  is a  $(p + 1) \times q$  matrix of regression coefficients ( $\mathbf{0}$  being the  $n \times q$  zero matrix). Hence, in principle, it is not unreasonable to assume  $n > p + 1$ ; in practice  $p$  will be much smaller than  $n$ , as we illustrate in the Discussion. Notice that the  $p$  predictors to be used will not be known *a priori*, and therefore it will be necessary to carry out variable selection together with covariance selection; this will be feasible using the marginal likelihoods corresponding to different design matrices. For simplicity, we will use a single  $\mathbf{X}$  in our notation (without explicitly conditioning on it).

In section 3.1 we summarize the main features of a standard conjugate analysis of the model

$$\mathbf{Y} \mid \mathbf{B}, \mathbf{\Omega} \sim \mathcal{N}_{n,q}(\mathbf{X}\mathbf{B}, \mathbf{I}_n, \mathbf{\Omega}^{-1}), \quad (11)$$

with  $\mathbf{B}$  unconstrained. This is done for completeness and for the benefit of the reader, so that the subsequent sections can be followed more easily; see also Rowe (2003), whose notation is somewhat different from ours. Next, in section 3.2, we derive the marginal data distribution for a subset of variables (selected columns of  $\mathbf{Y}$ ) which represents the building block for computing the marginal likelihood of a general DAG model (as detailed in section 5.1). We remark that, because of the theory presented in section 5.1, we need only consider an unconstrained  $\mathbf{\Omega}$  even when we deal with covariance matrices having a graphical structure. This is indeed a major simplification characterizing the approach taken in this paper; we will return to this issue later on.

### 3.1 Conjugate analysis

If we denote by  $\hat{\mathbf{B}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}$  the least squares estimator of  $\mathbf{B}$ , the likelihood function can be written as

$$f(\mathbf{Y} | \mathbf{B}, \Omega) = \frac{|\Omega|^{\frac{n}{2}}}{(2\pi)^{\frac{nq}{2}}} \exp \left\{ -\frac{1}{2} \text{tr} \left( \Omega \{ (\mathbf{B} - \hat{\mathbf{B}})^\top \mathbf{X}^\top \mathbf{X} (\mathbf{B} - \hat{\mathbf{B}}) + \hat{\mathbf{E}}^\top \hat{\mathbf{E}} \} \right) \right\}, \quad (12)$$

where  $\hat{\mathbf{E}} = (\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}})$  is the matrix of residuals. Hence, a conjugate prior for  $(\mathbf{B}, \Omega)$  is obtained by letting

$$\begin{aligned} \mathbf{B} | \Omega &\sim \mathcal{N}_{p+1,q}(\underline{\mathbf{B}}, \mathbf{C}^{-1}, \Omega^{-1}), \\ \Omega &\sim \mathcal{W}_q(a, \mathbf{R}), \end{aligned}$$

which results in the prior density

$$p(\mathbf{B}, \Omega) = \frac{|\Omega|^{\frac{(p+1)+(a-q-1)}{2}}}{K(\mathbf{C}, \mathbf{R}, a)} \exp \left\{ -\frac{1}{2} \text{tr} \left( \Omega \{ (\mathbf{B} - \underline{\mathbf{B}})^\top \mathbf{C} (\mathbf{B} - \underline{\mathbf{B}}) + \mathbf{R} \} \right) \right\}, \quad (13)$$

where

$$K(\mathbf{C}, \mathbf{R}, a) = \frac{(2\pi)^{\frac{q(p+1)}{2}} 2^{\frac{aq}{2}} \Gamma_q(\frac{a}{2})}{|\mathbf{C}|^{\frac{q}{2}} |\mathbf{R}|^{\frac{a}{2}}} \quad (14)$$

is the prior normalizing constant. The prior (13) is a matrix normal Wishart.

Some algebraic manipulations show that the posterior distribution of  $(\mathbf{B}, \Omega)$  is

$$\begin{aligned} \mathbf{B} | \Omega, \mathbf{Y} &\sim \mathcal{N}_{p+1,q}(\overline{\mathbf{B}}, (\mathbf{C} + \mathbf{X}^\top \mathbf{X})^{-1}, \Omega^{-1}), \\ \Omega | \mathbf{Y} &\sim \mathcal{W}_q(a + n, \mathbf{R} + \hat{\mathbf{E}}^\top \hat{\mathbf{E}} + \mathbf{D}), \end{aligned}$$

where  $\overline{\mathbf{B}} = (\mathbf{C} + \mathbf{X}^\top \mathbf{X})^{-1} (\mathbf{X}^\top \mathbf{Y} + \mathbf{C}\underline{\mathbf{B}})$  is the posterior expectation (matrix) of  $\mathbf{B}$ , and  $\mathbf{D} = (\underline{\mathbf{B}} - \hat{\mathbf{B}})^\top \{ \mathbf{C}^{-1} + (\mathbf{X}^\top \mathbf{X})^{-1} \}^{-1} (\underline{\mathbf{B}} - \hat{\mathbf{B}})$  is a measure of discrepancy between  $\underline{\mathbf{B}}$  and  $\hat{\mathbf{B}}$  (prior and data). Prior-to-posterior updating thus takes the form

$$\underline{\mathbf{B}} \mapsto \overline{\mathbf{B}}, \quad \mathbf{C} \mapsto \mathbf{C} + \mathbf{X}^\top \mathbf{X}, \quad a \mapsto a + n, \quad \mathbf{R} \mapsto \mathbf{R} + \hat{\mathbf{E}}^\top \hat{\mathbf{E}} + \mathbf{D}, \quad (15)$$

and the posterior density  $p(\mathbf{B}, \Omega | \mathbf{Y})$  is as in (13) with hyper-parameters updated by (15); the posterior normalizing constant will be given by

$$K(\mathbf{C} + \mathbf{X}^\top \mathbf{X}, \mathbf{R} + \hat{\mathbf{E}}^\top \hat{\mathbf{E}} + \mathbf{D}, a + n), \quad (16)$$

with the function  $K(\cdot, \cdot, \cdot)$  defined in (14).

### 3.2 Marginal data distribution

The marginal distribution of the matrix  $\mathbf{Y}$  can be obtained as

$$m(\mathbf{Y}) = \frac{f(\mathbf{Y} | \mathbf{B}, \boldsymbol{\Omega}) p(\mathbf{B}, \boldsymbol{\Omega})}{p(\mathbf{B}, \boldsymbol{\Omega} | \mathbf{Y})},$$

which in light of conjugacy gives

$$m(\mathbf{Y}) = \frac{K(\mathbf{C} + \mathbf{X}^\top \mathbf{X}, \mathbf{R} + \hat{\mathbf{E}}^\top \hat{\mathbf{E}} + \mathbf{D}, a + n)}{(2\pi)^{\frac{nq}{2}} K(\mathbf{C}, \mathbf{R}, a)}, \quad (17)$$

that is, up to a multiplicative factor, the ratio of the posterior and prior normalizing constants, (16) and (14), respectively.

In the sequel, we will also need the marginal distribution of selected columns of the data matrix  $\mathbf{Y}$ , corresponding to a proper subset  $J$  of the full set of  $q$  response variables. Let  $\mathbf{Y}_J$  be the  $n \times |J|$  selected data submatrix, and  $\mathbf{B}_J$  be the corresponding  $(p+1) \times |J|$  submatrix of  $\mathbf{B}$ , whose columns contain the regression coefficients for the selected responses. When restricted to the set  $J$  of response variables, by the results presented in section 2, the Gaussian multivariate regression model (11) can be written as

$$\mathbf{Y}_J | \mathbf{B}_J, \boldsymbol{\Omega}_{JJ \cdot \bar{J}} \sim \mathcal{N}_{n, |J|}(\mathbf{X} \mathbf{B}_J, \mathbf{I}_n, \boldsymbol{\Omega}_{JJ \cdot \bar{J}}^{-1}),$$

with induced prior

$$\begin{aligned} \mathbf{B}_J | \boldsymbol{\Omega}_{JJ \cdot \bar{J}} &\sim \mathcal{N}_{p+1, |J|}(\underline{\mathbf{B}}_J, \mathbf{C}^{-1}, \boldsymbol{\Omega}_{JJ \cdot \bar{J}}^{-1}), \\ \boldsymbol{\Omega}_{JJ \cdot \bar{J}} &\sim \mathcal{W}_{|J|}(a - |\bar{J}|, \mathbf{R}_{JJ}), \end{aligned}$$

where  $\underline{\mathbf{B}}_J$  is the appropriate submatrix of  $\underline{\mathbf{B}}$ .

One readily sees that the formal structure of model and prior for a subset  $J$  of response variables is the same as for the full data matrix. As a consequence, the marginal data distribution for the submatrix  $\mathbf{Y}_J$  is given by (17) with the following substitutions:

$$q \mapsto |J|, \quad \mathbf{R} \mapsto \mathbf{R}_{JJ}, \quad a \mapsto a - |\bar{J}|, \quad \underline{\mathbf{B}} \mapsto \underline{\mathbf{B}}_J, \quad \hat{\mathbf{B}} \mapsto \hat{\mathbf{B}}_J, \quad \hat{\mathbf{E}} \mapsto \hat{\mathbf{E}}_J, \quad \mathbf{D} \mapsto \mathbf{D}_{JJ},$$

while  $n$ ,  $\mathbf{C}$  and  $\mathbf{X}$  remain unchanged.

## 4 Objective analysis

We assume the reader is familiar with the basic concepts of model selection from the Bayesian perspective, as described for instance in O’Hagan & Forster (2004, ch. 7). Here, in section 4.1, we provide some background on *objective Bayes* model selection, focusing in particular on a proposal by O’Hagan (1995). Then, in section 4.2, we give the expression for the marginal data distribution of a generic subset of columns of  $\mathbf{Y}$  under the prior implied by such proposal; this will be instrumental in the construction of the marginal likelihood of a DAG model given in section 5.1.

### 4.1 Fractional parameter priors

Let  $\mathcal{M}_1, \dots, \mathcal{M}_K$  be a collection of Bayesian models for the same observable  $\mathbf{Y}$ . Each model  $\mathcal{M}_k$ ,  $k = 1, \dots, K$ , consists of a family of sampling densities  $f_k(\mathbf{Y} \mid \boldsymbol{\theta}_k)$ , indexed by a model specific parameter  $\boldsymbol{\theta}_k$ , and of a prior density  $p_k(\boldsymbol{\theta}_k)$  on  $\boldsymbol{\theta}_k$ , which we assume to be *proper*. We focus on the comparison of  $\mathcal{M}_k$  with  $\mathcal{M}_{k'}$  through the Bayes factor. The Bayes factor for  $\mathcal{M}_k$  against  $\mathcal{M}_{k'}$  is defined as  $BF_{kk'}(\mathbf{Y}) = m_k(\mathbf{Y})/m_{k'}(\mathbf{Y})$ , where  $m_k(\mathbf{Y}) = \int f_k(\mathbf{Y} \mid \boldsymbol{\theta}_k)p_k(\boldsymbol{\theta}_k)d\boldsymbol{\theta}_k$  is the marginal density of  $\mathbf{Y}$  under  $\mathcal{M}_k$ , also known as the marginal likelihood of  $\mathcal{M}_k$ .

In lack of substantive prior information, we would like to take  $p_k(\boldsymbol{\theta}_k) = p_k^D(\boldsymbol{\theta}_k)$  for some objective default (non-informative) parameter prior  $p_k^D(\boldsymbol{\theta}_k)$ . However, objective priors are often improper and they cannot be naively used to compute Bayes factors, even when the marginal likelihoods  $m_k(\mathbf{Y})$  are finite and non-zero, because of the presence of arbitrary constants which do not cancel out in their ratios. Pericchi (2005) reviews several proposals put forward to address this issue. In this paper, we take advantage of the fractional Bayes factor originally introduced by O’Hagan (1995); see also O’Hagan & Forster (2004).

Let  $b = b(n)$ ,  $0 < b < 1$ , be a fraction of the number of observations  $n$ . Define

$$m_k(\mathbf{Y}; b) = \frac{\int f_k(\mathbf{Y} \mid \boldsymbol{\theta}_k)p_k^D(\boldsymbol{\theta}_k)d\boldsymbol{\theta}_k}{\int f_k^b(\mathbf{Y} \mid \boldsymbol{\theta}_k)p_k^D(\boldsymbol{\theta}_k)d\boldsymbol{\theta}_k}, \quad (18)$$

where  $f_k^b(\mathbf{Y} \mid \boldsymbol{\theta}_k)$  is the sampling density under model  $\mathcal{M}_k$  raised to the  $b$ -th power,

and the two integrals are assumed to be finite and non-zero. The *fractional marginal likelihood* (18) of model  $\mathcal{M}_k$ , can be rewritten as

$$m_k(\mathbf{Y}; b) = \int f_k^{1-b}(\mathbf{Y} | \boldsymbol{\theta}_k) p_k^F(\boldsymbol{\theta}_k | b, \mathbf{Y}) d\boldsymbol{\theta}_k,$$

where  $p_k^F(\boldsymbol{\theta}_k | b, \mathbf{Y}) \propto f_k^b(\mathbf{Y} | \boldsymbol{\theta}_k) p_k^D(\boldsymbol{\theta}_k)$  is the implied *fractional prior* (actually a “posterior” based on the fractional likelihood and the default prior). The fractional Bayes factor for  $\mathcal{M}_k$  against  $\mathcal{M}_{k'}$  is then defined as the ratio of  $m_k(\mathbf{Y}; b)$  to  $m_{k'}(\mathbf{Y}; b)$ . In essence, a fraction of the data is used to obtain a proper prior, which is then applied to the remaining fraction.

Clearly, the fractional prior depends on the choice of  $b$ . Usually  $b$  will be small, so that dependence of the prior on the data will be weak. Consistency is achieved as long as  $b \rightarrow 0$  for  $n \rightarrow \infty$ . O’Hagan (1995, sect. 4) suggests  $b = n_0/n$  as a default choice, where  $n_0$  is the minimal (integer) training sample size for which the fractional marginal likelihood is well defined, together with a couple of alternative choices, to be used when robustness is an issue. Moreno (1997) has an argument according to which the default choice is the only valid one, and we stick to this choice in this paper.

## 4.2 Fractional marginal likelihoods

Consider the Gaussian multivariate regression model (11). We start from the prior

$$p^D(\mathbf{B}, \boldsymbol{\Omega}) \propto |\boldsymbol{\Omega}|^{\frac{a_D - q - 1}{2}}, \quad (19)$$

$\boldsymbol{\Omega}$  s.p.d., which is flexible enough to accommodate different choices of default priors. In particular,  $a_D = 0$  gives  $p^D(\mathbf{B}, \boldsymbol{\Omega}) \propto |\boldsymbol{\Omega}|^{\frac{-(q+1)}{2}}$ , equivalently  $p^D(\mathbf{B}, \boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{\frac{-(q+1)}{2}}$  for  $\boldsymbol{\Sigma} = \boldsymbol{\Omega}^{-1}$ , because the Jacobian of  $(\mathbf{B}, \boldsymbol{\Omega}) \mapsto (\mathbf{B}, \boldsymbol{\Sigma})$  is proportional to  $|\boldsymbol{\Sigma}|^{(q+1)}$ . This is the “independence” Jeffreys prior, that is, the prior obtained by multiplying the Jeffreys priors for the two parameters assuming the other one is known; see Press (1982, sect. 3.6.2 and (14.2.7)). Alternatively,  $a_D = q - 1$  gives  $p^D(\mathbf{B}, \boldsymbol{\Omega}) \propto |\boldsymbol{\Omega}|^{-1}$ , or  $p^D(\mathbf{B}, \boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-q}$ . Both these priors are discussed in Geisser & Cornfield (1963), whereas Geisser (1965) focusses more deeply on the independence Jeffreys prior. Sun & Berger (2007) present further objective priors for the multivariate normal model.

Using the default prior (19), and setting the fraction  $b$  equal to  $n_0/n$ , the fractional prior for the multivariate regression model (11) is given by

$$p(\mathbf{B}, \mathbf{\Omega}) \propto |\mathbf{\Omega}|^{\frac{a_D + n_0 - p - q - 2}{2}} \exp \left\{ -\frac{n_0}{2} \text{tr} \left( \mathbf{\Omega} \left\{ (\mathbf{B} - \hat{\mathbf{B}})^\top \tilde{\mathbf{C}} (\mathbf{B} - \hat{\mathbf{B}}) + \tilde{\mathbf{R}} \right\} \right) \right\}, \quad (20)$$

where  $\tilde{\mathbf{C}} = n^{-1} \mathbf{X}^\top \mathbf{X}$  and  $\tilde{\mathbf{R}} = n^{-1} \hat{\mathbf{E}}^\top \hat{\mathbf{E}}$ ; this is clearly a matrix normal Wishart, having the form (13) with

$$\underline{\mathbf{B}} = \hat{\mathbf{B}}, \quad \mathbf{C} = n_0 \tilde{\mathbf{C}}, \quad a = a_D + n_0 - p - 1, \quad \mathbf{R} = n_0 \tilde{\mathbf{R}}.$$

The prior (20) is proper under two conditions: i)  $a_D + n_0 - p > q$ , so that  $a > q - 1$ ; ii)  $n - p - 1 > q - 1$ , so that  $\hat{\mathbf{E}}^\top \hat{\mathbf{E}}$  is (almost surely) positive definite.

Condition ii), which simplifies to  $n > p + q$ , will not be met in our intended application setting, but we will be able to relax it in the context of sparse DAG models; see section 5.1. Condition i) becomes  $n_0 > p + q$ , if  $a_D = 0$ , or  $n_0 > p + 1$ , if  $a_D = q - 1$ . Clearly, the fraction  $b = n_0/n$  must be larger when using the independence Jeffreys prior, rather than the prior presented in Geisser & Cornfield (1963), especially if  $q$  is much larger than 1. Since the fraction of the data to be used should be as small as possible, we recommend setting  $a_D = q - 1$  (and  $n_0 = p + 2$ , so that  $a = q$ ). Notice that, for  $b = n_0/n$  to be small, with  $n_0 > p + 1$ , we need  $p \ll n$ , which is a stronger requirement than assuming  $n > p + 1$  as in section 3. However, as anticipated in section 3, and illustrated in the Discussion, this requirement will be typically satisfied in our intended application setting.

Posterior updating of the hyper-parameters leads to

$$\overline{\mathbf{B}} = \hat{\mathbf{B}}, \quad \mathbf{C} \mapsto n \tilde{\mathbf{C}}, \quad a \mapsto a_D + n - p - 1, \quad \mathbf{R} \mapsto n \tilde{\mathbf{R}},$$

keeping into account that the fractional prior is to be used on the likelihood raised to the  $(1 - b)$ -th power, that is, on data with the same  $\hat{\mathbf{B}}$ ,  $\tilde{\mathbf{C}}$  and  $\tilde{\mathbf{R}}$ , but with  $n - n_0$  in place of  $n$ . Consequently, using (17), one gets

$$m(\mathbf{Y}) = \frac{K(\mathbf{X}^\top \mathbf{X}, \hat{\mathbf{E}}^\top \hat{\mathbf{E}}, a_D + n - p - 1)}{(2\pi)^{\frac{nq}{2}} K(n_0 n^{-1} \hat{\mathbf{X}}^\top \hat{\mathbf{X}}, n_0 n^{-1} \hat{\mathbf{E}}^\top \hat{\mathbf{E}}, a_D + n_0 - p - 1)},$$

which after some simplifications leads to

$$m(\mathbf{Y}) = \pi^{-\frac{(n-n_0)q}{2}} \frac{\Gamma_q\left(\frac{a_D + n - p - 1}{2}\right)}{\Gamma_q\left(\frac{a_D + n_0 - p - 1}{2}\right)} \left(\frac{n_0}{n}\right)^{\frac{q(a_D + n_0)}{2}} |\hat{\mathbf{E}}^\top \hat{\mathbf{E}}|^{-\frac{n-n_0}{2}}. \quad (21)$$

In order to apply the method presented in section 5 one also needs the fractional marginal likelihood based on the submatrix  $\mathbf{Y}_J$  which only contains the columns of  $\mathbf{Y}$  belonging to the subset  $J$ , which we write as  $m(\mathbf{Y}_J)$ . This marginal likelihood is germane to our approach, and represents a half-way house towards computing the entire fractional marginal likelihood for a DAG model; see section 5.1. Based on the results presented in section 3.2, it is immediate to conclude that  $m(\mathbf{Y}_J)$  can be obtained from equation (21) upon making the substitutions

$$q \mapsto |J|, \quad a_D \mapsto a_D - |\bar{J}|, \quad \hat{\mathbf{E}} \mapsto \hat{\mathbf{E}}_J = (\mathbf{Y}_J - \mathbf{X}\hat{\mathbf{B}}_J).$$

These substitutions lead to

$$m(\mathbf{Y}_J) = \pi^{-\frac{(n-n_0)|J|}{2}} \frac{\Gamma_{|J|}\left(\frac{a_D+n-p-1-|\bar{J}|}{2}\right)}{\Gamma_{|J|}\left(\frac{a_D+n_0-p-1-|\bar{J}|}{2}\right)} \left(\frac{n_0}{n}\right)^{\frac{|J|(a_D+n_0-|\bar{J}|)}{2}} |\hat{\mathbf{E}}_J^\top \hat{\mathbf{E}}_J|^{-\frac{n-n_0}{2}}, \quad (22)$$

which returns (21) upon setting  $J = \{1, \dots, q\}$ .

Formula (22) derives from  $\boldsymbol{\Omega}_{JJ, \bar{J}} \sim \mathcal{W}_{|J|}(a_J, \mathbf{R}_{JJ})$  with  $a_J = a_D + n_0 - p - 1 - |\bar{J}|$ , which is (almost surely) proper if  $n > p + |J|$ . The latter condition guarantees positive definiteness of  $\mathbf{R}_{JJ}$ , while  $a_J = q - |\bar{J}| = |J|$  using our recommended choices for  $a_D$  and  $n_0$ . Therefore, formula (22) provides us with a valid value for  $m(\mathbf{Y}_J)$ , whenever  $|J| < n - p$ , even if  $n \leq p + q$ . We will exploit this fact in section 5.1 to derive the marginal likelihood of a sparse DAG. In the next paragraph we specialize (22) to the simplest regression setup, which is of some interest in its own right.

If the sampling distribution corresponds to i.i.d. observations from a  $q$ -dimensional Gaussian density with expectation  $\boldsymbol{\mu}$  and precision  $\boldsymbol{\Omega}$ , conditionally on  $\boldsymbol{\mu}$  and  $\boldsymbol{\Omega}$ , the corresponding marginal data distribution  $m(\mathbf{Y}_J)$  can be derived from (22) upon setting  $p = 0$  (no predictors) and  $\hat{\mathbf{E}} = \mathbf{Y} - \mathbf{1}_n \bar{\mathbf{y}}^\top$ , where  $\bar{\mathbf{y}}$  is the  $q$ -dimensional vector of sample means. In this way we obtain

$$m(\mathbf{Y}_J) = \pi^{-\frac{(n-n_0)|J|}{2}} \frac{\Gamma_{|J|}\left(\frac{a_D+n-1-|\bar{J}|}{2}\right)}{\Gamma_{|J|}\left(\frac{a_D+n_0-1-|\bar{J}|}{2}\right)} \left(\frac{n_0}{n}\right)^{\frac{|J|(a_D+n_0-|\bar{J}|)}{2}} |\hat{\mathbf{E}}_J^\top \hat{\mathbf{E}}_J|^{-\frac{n-n_0}{2}}, \quad (23)$$

with  $(\hat{\mathbf{E}}^\top \hat{\mathbf{E}})_{jj'} = \sum_i (y_{ij} - \bar{y}_j)(y_{ij'} - \bar{y}_{j'})$ . Expression (23) complements formula (22) in Consonni & La Rocca (2012), which holds for i.i.d.  $q$ -dimensional Gaussian observations with zero expectation.



## 5 Covariance selection

So far we have analyzed the Gaussian multivariate regression model (11) under the condition that  $\mathbf{\Omega}$  is unconstrained. We now assume instead that  $\mathbf{\Omega}$  is constrained by a DAG, aiming at graphical model (or covariance) selection after having adjusted for the presence of covariates. In section 5.1, we develop an extension of the approach by Geiger & Heckerman (2002) explicitly for the regression setup. An advantage of the method we present is that the computation of the marginal likelihood for each DAG only requires the results established, for an unconstrained  $\mathbf{\Omega}$ , in section 4.2. In section 5.2, taking advantage of the fact that any two Markov equivalent DAGs obtain the same marginal likelihood, we specify our results to the case of Gaussian decomposable graphical models, and relate them to those obtained by Carvalho & Scott (2009) in the i.i.d. case.

### 5.1 Acyclic directed error structure

Let  $\mathcal{D}$  be a DAG with vertex set  $\{1, \dots, q\}$ . Denote by  $\text{pa}_{\mathcal{D}}(j)$  the *parents* of  $j$  in  $\mathcal{D}$ , that is, the set of all vertices in  $\mathcal{D}$  from which an edge points to vertex  $j$ , and by  $\mathbf{y}_{\text{ipa}_{\mathcal{D}}(j)}$  the subvector of  $\mathbf{y}_i$  indexed by  $\text{pa}_{\mathcal{D}}(j)$ . The multivariate normal sampling density of  $\mathbf{y}_i | \mathbf{B}, \mathbf{\Omega}$ , assumed to be Markov with respect to  $\mathcal{D}$ , can be written as

$$f_{\mathcal{D}}(\mathbf{y}_i | \boldsymbol{\theta}_{\mathcal{D}}) = \prod_{j=1}^q f_{\mathcal{D}}(y_{ij} | \mathbf{y}_{\text{ipa}_{\mathcal{D}}(j)}; \boldsymbol{\theta}_j), \quad (24)$$

where  $\boldsymbol{\theta}_j = (\boldsymbol{\alpha}_j, \boldsymbol{\gamma}_j, \lambda_j)$  is defined by

$$\mathbb{E}(y_{ij} | \mathbf{y}_{\text{ipa}_{\mathcal{D}}(j)}; \mathbf{B}, \mathbf{\Omega}) = \mathbf{x}_i^{\top} \boldsymbol{\alpha}_j + \mathbf{y}_{\text{ipa}_{\mathcal{D}}(j)}^{\top} \boldsymbol{\gamma}_j, \quad (25)$$

$$\text{Var}(y_{ij} | \mathbf{y}_{\text{ipa}_{\mathcal{D}}(j)}; \mathbf{B}, \mathbf{\Omega}) = \lambda_j^{-1}, \quad (26)$$

and  $\boldsymbol{\theta}_{\mathcal{D}} = (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_q)$  is the collection of all  $\boldsymbol{\theta}_j$ s; recall that  $\mathbf{x}_i^{\top}$  is the  $i$ -th row of the design matrix  $\mathbf{X}$ , and notice that we drop dependence on  $\mathcal{D}$  when we move from  $\boldsymbol{\theta}_{\mathcal{D}}$  to its components (to lighten notation). We illustrate below the reparameterization from  $(\mathbf{B}, \mathbf{\Omega})$ , with  $\mathbf{\Omega}$  s.p.d., to  $\boldsymbol{\theta}_{\mathcal{D}}$ , with  $\lambda_j > 0$ ,  $j = 1, \dots, q$ , after a remark on (24).

The conditional vertex density  $f_{\mathcal{D}}(y_{ij} | \mathbf{y}_{i\text{pa}_{\mathcal{D}}(j)}; \boldsymbol{\theta}_j)$  is a univariate normal density with expectation and variance given by (25) and (26), respectively. It is important to remark that such density depends on  $\mathcal{D}$  only through  $\text{pa}_{\mathcal{D}}(j)$ . In other words, if two DAGs  $\mathcal{D}_1$  and  $\mathcal{D}_2$  are such that  $\text{pa}_{\mathcal{D}_1}(j) = \text{pa}_{\mathcal{D}_2}(j)$ , then the vertex-specific parameter  $\boldsymbol{\theta}_j$  varies in the same space under  $\mathcal{D}_1$  and  $\mathcal{D}_2$ , because  $\boldsymbol{\gamma}_j$  has the same dimension under the two DAGs, and  $f_{\mathcal{D}_1}(y_{ij} | \mathbf{y}_{i\text{pa}_{\mathcal{D}_1}(j)}; \boldsymbol{\theta}_j) = f_{\mathcal{D}_2}(y_{ij} | \mathbf{y}_{i\text{pa}_{\mathcal{D}_2}(j)}; \boldsymbol{\theta}_j)$ . This property, called *likelihood modularity* by Geiger & Heckerman (2002), represents a condition to be satisfied for the subsequent theory to apply.

Assume (without loss of generality) that the vertices of  $\mathcal{D}$  are well-numbered; this means that, if  $j'$  is a parent of  $j$ , then  $j' < j$ . If  $\mathcal{D}$  is *complete*, that is, it has all pairs of vertices joined by an edge, then the parameters indexing the last ( $j = q$ ) conditional vertex density in (24) are:  $\boldsymbol{\alpha}_q = \mathbf{B}_q + \mathbf{B}_{\bar{q}}\boldsymbol{\Omega}_{\bar{q}q}\boldsymbol{\Omega}_{qq}^{-1}$ ,  $\boldsymbol{\gamma}_q = -\boldsymbol{\Omega}_{\bar{q}q}\boldsymbol{\Omega}_{qq}^{-1}$ , and  $\lambda_q = \Omega_{qq}$ , where  $\bar{q} = \{1, \dots, q-1\} = \text{pa}_{\mathcal{D}}(q)$ ; see the end of section 2.1. Then, since  $\mathbf{y}_{i\bar{q}} | \mathbf{B}, \boldsymbol{\Omega} \sim \mathcal{N}_{q-1}(\mathbf{B}_{\bar{q}}^\top \mathbf{x}_i, \boldsymbol{\Omega}_{\bar{q}\bar{q}.q}^{-1})$ , one can repeat the previous argument and recursively find  $\boldsymbol{\theta}_{q-1}, \dots, \boldsymbol{\theta}_1$ . If  $\mathcal{D}$  is *incomplete*, its missing edges will impose on  $\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_q$  the constraints  $\gamma_{jj'} = 0$ ,  $j' \notin \text{pa}_{\mathcal{D}}(j)$ ,  $j = 1, \dots, q$ , so that a corresponding set of constraints will be imposed on  $\boldsymbol{\Omega}$ .

We now show that, for complete DAGs, the transformation  $(\mathbf{B}, \boldsymbol{\Omega}) \mapsto \boldsymbol{\theta}_{\mathcal{D}}$  is a smooth bijection. This fact, which is arguably not new, is reported here because it will be used below for constructing priors under general DAGs. Given the recursive definition of  $(\mathbf{B}, \boldsymbol{\Omega}) \mapsto \boldsymbol{\theta}_{\mathcal{D}}$ , it is enough to show that the transformation from  $(\mathbf{B}, \boldsymbol{\Omega})$ , with  $\boldsymbol{\Omega}$  s.p.d., to  $(\mathbf{B}_{\bar{q}}, \boldsymbol{\Omega}_{\bar{q}\bar{q}.q}; \boldsymbol{\alpha}_q, \boldsymbol{\gamma}_q, \lambda_q)$ , with  $\boldsymbol{\Omega}_{\bar{q}\bar{q}.q}$  s.p.d. and  $\lambda_q > 0$ , is a smooth bijection. We do this by composing a few simpler reparameterizations. First, we go from  $(\mathbf{B}, \boldsymbol{\Omega})$ , with  $\boldsymbol{\Omega}$  s.p.d., to  $(\mathbf{B}, \boldsymbol{\Omega}_{\bar{q}\bar{q}.q}, \boldsymbol{\Omega}_{\bar{q}q}, \Omega_{qq})$ , with  $\boldsymbol{\Omega}_{\bar{q}\bar{q}.q}$  s.p.d. and  $\Omega_{qq} > 0$ , where the smooth inverse map is provided by  $\boldsymbol{\Omega}_{\bar{q}\bar{q}} = \boldsymbol{\Omega}_{\bar{q}\bar{q}.q} + \boldsymbol{\Omega}_{\bar{q}q}\boldsymbol{\Omega}_{qq}^{-1}\boldsymbol{\Omega}_{\bar{q}q}^\top$ , recalling that  $\boldsymbol{\Omega}_{q\bar{q}} = \boldsymbol{\Omega}_{\bar{q}q}^\top$  (unconstrained); see for instance Lauritzen (1996, Lemma B.1). Then, we trivially split  $\mathbf{B}$  as  $(\mathbf{B}_q, \mathbf{B}_{\bar{q}})$ , and replace  $\mathbf{B}_q$  with  $\boldsymbol{\alpha}_q$ , where the smooth inverse map is given by  $\mathbf{B}_q = \boldsymbol{\alpha}_q - \mathbf{B}_{\bar{q}}\boldsymbol{\Omega}_{\bar{q}q}\boldsymbol{\Omega}_{qq}^{-1}$ . Finally, we reparameterize from  $\boldsymbol{\Omega}_{\bar{q}q}$  to  $\boldsymbol{\gamma}_q$ , with smooth inverse map given by  $\boldsymbol{\Omega}_{\bar{q}q} = -\boldsymbol{\Omega}_{qq}\boldsymbol{\gamma}_q$ , and we rename  $\Omega_{qq}$  as  $\lambda_q$  (constrained to be positive).

In light of the above discussion, all complete DAGs define the same statistical model, in which  $\mathbf{\Omega}$  is unconstrained, and there is a smooth bijection between their collections of parameters; in the terminology of Geiger & Heckerman (2002) we have *complete model equivalence*, and *regularity*. It follows that any prior on  $(\mathbf{B}, \mathbf{\Omega})$  will induce a prior on  $\boldsymbol{\theta}_{\mathcal{D}}$ , if  $\mathcal{D}$  is complete. We now show that, if we let  $(\mathbf{B}, \mathbf{\Omega})$  follow the conjugate prior (13), then  $p_{\mathcal{D}}(\boldsymbol{\theta}_{\mathcal{D}}) = \prod_{j=1}^q p_{\mathcal{D}}(\boldsymbol{\theta}_j)$ , so that  $\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_q$  will be *a priori* independent. This property is called *global parameter independence*, and represents a crucial ingredient in the approach of Geiger & Heckerman (2002); it can be obtained by recursive application of the following result.

**Proposition 5.1.** *If  $\mathbf{B} \mid \mathbf{\Omega} \sim \mathcal{N}_{(p+1) \times q}(\underline{\mathbf{B}}, \mathbf{C}^{-1}, \mathbf{\Omega}^{-1})$  and  $\mathbf{\Omega} \sim \mathcal{W}_q(a, \mathbf{R})$ , then the pair  $(\mathbf{B}_{\bar{q}}, \mathbf{\Omega}_{\bar{q}\bar{q}\cdot q})$  is independent of the triple  $(\mathbf{B}_q + \mathbf{B}_{\bar{q}}\mathbf{\Omega}_{\bar{q}q}\mathbf{\Omega}_{qq}^{-1}, \mathbf{\Omega}_{\bar{q}q}, \mathbf{\Omega}_{qq})$ .*

*Proof.* Consider the reparameterization in terms of  $\mathbf{\Omega}_{\bar{q}\bar{q}\cdot q}$  s.p.d.,  $\mathbf{\Omega}_{\bar{q}q}, \mathbf{\Omega}_{qq} > 0$ ,  $\mathbf{B}_{\bar{q}}$ ,  $\boldsymbol{\alpha}_q = \mathbf{B}_q + \mathbf{B}_{\bar{q}}\mathbf{\Omega}_{\bar{q}q}\mathbf{\Omega}_{qq}^{-1}$ , and factorize the corresponding joint parameter density as

$$p(\boldsymbol{\alpha}_q \mid \mathbf{B}_{\bar{q}}, \mathbf{\Omega}_{\bar{q}\bar{q}\cdot q}, \mathbf{\Omega}_{\bar{q}q}, \mathbf{\Omega}_{qq}) \times p(\mathbf{B}_{\bar{q}} \mid \mathbf{\Omega}_{\bar{q}\bar{q}\cdot q}, \mathbf{\Omega}_{\bar{q}q}, \mathbf{\Omega}_{qq}) \times p(\mathbf{\Omega}_{\bar{q}\bar{q}\cdot q}, \mathbf{\Omega}_{\bar{q}q}, \mathbf{\Omega}_{qq}).$$

We know, from our statement following (10), that  $\mathbf{\Omega}_{\bar{q}\bar{q}\cdot q}$  is independent of  $(\mathbf{\Omega}_{\bar{q}q}, \mathbf{\Omega}_{qq})$  under the assumed distribution for  $\mathbf{\Omega}$ . Moreover, from the law of  $\mathbf{B} \mid \mathbf{\Omega}$ , we obtain

$$\begin{aligned} \mathbf{B}_{\bar{q}} \mid \mathbf{\Omega}_{\bar{q}\bar{q}\cdot q}, \mathbf{\Omega}_{\bar{q}q}, \mathbf{\Omega}_{qq} &\sim \mathcal{N}_{(p+1), (q-1)}(\underline{\mathbf{B}}_{\bar{q}}, \mathbf{C}^{-1}, \mathbf{\Omega}_{\bar{q}\bar{q}\cdot q}^{-1}), \\ \boldsymbol{\alpha}_q \mid \mathbf{B}_{\bar{q}}, \mathbf{\Omega}_{\bar{q}\bar{q}\cdot q}, \mathbf{\Omega}_{\bar{q}q}, \mathbf{\Omega}_{qq} &\sim \mathcal{N}_{p+1}(\underline{\mathbf{B}}_q - \underline{\mathbf{B}}_{\bar{q}}\mathbf{\Omega}_{\bar{q}q}\mathbf{\Omega}_{qq}^{-1}, \mathbf{\Omega}_{qq}^{-1}\mathbf{C}^{-1}), \end{aligned}$$

first using column marginalization (4), and (9), then using column conditioning (5).

Therefore, the joint density of  $\mathbf{\Omega}_{\bar{q}\bar{q}\cdot q}, \mathbf{\Omega}_{\bar{q}q}, \mathbf{\Omega}_{qq}, \mathbf{B}_{\bar{q}}$ , and  $\boldsymbol{\alpha}_q$ , factorizes as

$$p(\boldsymbol{\alpha}_q \mid \mathbf{\Omega}_{\bar{q}q}, \mathbf{\Omega}_{qq}) \times p(\mathbf{B}_{\bar{q}} \mid \mathbf{\Omega}_{\bar{q}\bar{q}\cdot q}) \times p(\mathbf{\Omega}_{\bar{q}\bar{q}\cdot q}) \times p(\mathbf{\Omega}_{\bar{q}q}, \mathbf{\Omega}_{qq}),$$

which implies the desired result.  $\square$

If  $\mathcal{D}$  is incomplete, global parameter independence can be guaranteed by letting  $p_{\mathcal{D}}(\boldsymbol{\theta}_{\mathcal{D}}) = \prod_{j=1}^q p_{\mathcal{C}_j}(\boldsymbol{\theta}_j)$ , where  $\mathcal{C}_j$  is any complete DAG such that  $\text{pa}_{\mathcal{C}_j}(j) = \text{pa}_{\mathcal{D}}(j)$ . The actual choice of each  $\mathcal{C}_j$  is immaterial, because all  $j' \notin \text{pa}_{\mathcal{D}}(j)$ ,  $j' \neq j$ , must follow  $j$  in  $\mathcal{C}_j$ , and thus  $p_{\mathcal{C}_j}(\boldsymbol{\theta}_j)$  is induced by the law of  $(\mathbf{B}_F, \mathbf{\Omega}_{F\bar{F}})$ , where  $F = \text{fa}_{\mathcal{D}}(j) =$

$\text{pa}_{\mathcal{D}}(j) \cup \{j\}$  is the *family* of  $j$  in  $\mathcal{D}$ . Notice that  $j$  goes necessarily last in  $\text{fa}_{\mathcal{D}}(j)$ , and recall that  $\mathbf{B}_F | \boldsymbol{\Omega}_{FF \cdot \bar{F}} \sim \mathcal{N}_{(p+1) \times |F|}(\underline{\mathbf{B}}_F, \mathbf{C}^{-1}, \boldsymbol{\Omega}_{FF \cdot \bar{F}}^{-1})$ , by column marginalization, while  $\boldsymbol{\Omega}_{FF \cdot \bar{F}} \sim \mathcal{W}_{|F|}(a - |F|, \mathbf{R}_{FF})$ , as per (10). Assigning parameter priors in this way, we also guarantee *prior modularity*:  $p_{\mathcal{D}_1}(\boldsymbol{\theta}_j) = p_{\mathcal{D}_2}(\boldsymbol{\theta}_j)$ , if  $\text{pa}_{\mathcal{D}_1}(j) = \text{pa}_{\mathcal{D}_2}(j)$ . This is the last ingredient required by the method of Geiger & Heckerman (2002) to compute the marginal likelihood of *any* DAG model, based on the assignment of the *single* prior (13). We now detail the computations for our regression setting.

The marginal density of the matrix  $\mathbf{Y}$  under the DAG  $\mathcal{D}$ , equivalently the marginal likelihood of  $\mathcal{D}$  observing  $\mathbf{Y}$ , can be found as  $m_{\mathcal{D}}(\mathbf{Y}) = \int f_{\mathcal{D}}(\mathbf{Y} | \boldsymbol{\theta}_{\mathcal{D}}) p_{\mathcal{D}}(\boldsymbol{\theta}_{\mathcal{D}}) d\boldsymbol{\theta}_{\mathcal{D}}$ , where  $f_{\mathcal{D}}(\mathbf{Y} | \boldsymbol{\theta}_{\mathcal{D}}) = \prod_{i=1}^n f_{\mathcal{D}}(\mathbf{y}_i | \boldsymbol{\theta}_{\mathcal{D}})$  with  $f_{\mathcal{D}}(\mathbf{y}_i | \boldsymbol{\theta}_{\mathcal{D}})$  given by (24), and furthermore  $p_{\mathcal{D}}(\boldsymbol{\theta}_{\mathcal{D}}) = \prod_{j=1}^q p_{\mathcal{D}}(\boldsymbol{\theta}_j)$  by global parameter independence. We can thus write

$$\begin{aligned} m_{\mathcal{D}}(\mathbf{Y}) &= \prod_{j=1}^q \int p_{\mathcal{D}}(\boldsymbol{\theta}_j) \prod_{i=1}^n f_{\mathcal{D}}(y_{ij} | \mathbf{y}_{i\text{pa}_{\mathcal{D}}(j)}; \boldsymbol{\theta}_j) d\boldsymbol{\theta}_j \\ &= \prod_{j=1}^q \int p_{\mathcal{C}_j}(\boldsymbol{\theta}_j) \prod_{i=1}^n f_{\mathcal{C}_j}(y_{ij} | \mathbf{y}_{i\text{pa}_{\mathcal{C}_j}(j)}; \boldsymbol{\theta}_j) d\boldsymbol{\theta}_j \\ &= \prod_{j=1}^q \int p_{\mathcal{C}_j}(\boldsymbol{\theta}_j) f_{\mathcal{C}_j}(\mathbf{Y}_j | \mathbf{Y}_{\text{pa}_{\mathcal{C}_j}(j)}; \boldsymbol{\theta}_j) d\boldsymbol{\theta}_j, \end{aligned}$$

where the second equality is based on prior and likelihood modularity. It follows that

$$m_{\mathcal{D}}(\mathbf{Y}) = \prod_{j=1}^q m_{\mathcal{C}_j}(\mathbf{Y}_j | \mathbf{Y}_{\text{pa}_{\mathcal{C}_j}(j)}) = \prod_{j=1}^q \frac{m_{\mathcal{C}_j}(\mathbf{Y}_{\text{fa}_{\mathcal{C}_j}(j)})}{m_{\mathcal{C}_j}(\mathbf{Y}_{\text{pa}_{\mathcal{C}_j}(j)})} = \prod_{j=1}^q \frac{m(\mathbf{Y}_{\text{fa}_{\mathcal{D}}(j)})}{m(\mathbf{Y}_{\text{pa}_{\mathcal{D}}(j)})}, \quad (27)$$

recalling that  $\text{pa}_{\mathcal{C}_j}(j) \equiv \text{pa}_{\mathcal{D}}(j)$ , by construction, and  $m_{\mathcal{C}_j}(\cdot)$  is nothing else but  $m(\cdot)$  under our prior (13), by complete model equivalence and regularity.

The great advantage of (27) is that the computations of the required terms in the rightmost product can be done under the assumption that the precision matrix  $\boldsymbol{\Omega}$  is unconstrained, and thus one can use the standard matrix normal Wishart prior (13). Notice that the DAG  $\mathcal{D}$  enters (27) only through the specification of the set of parents,  $\text{pa}_{\mathcal{D}}(j)$ , for each vertex  $j$ . The expressions for  $m(\mathbf{Y}_{\text{fa}_{\mathcal{D}}(j)})$  and  $m(\mathbf{Y}_{\text{pa}_{\mathcal{D}}(j)})$  are available in section 3.2, upon replacing  $J$  with  $\text{fa}_{\mathcal{D}}(j)$  and  $\text{pa}_{\mathcal{D}}(j)$ , respectively.

Prior (13) requires to specify the hyper-parameters  $\underline{\mathbf{B}}$ ,  $\mathbf{C}$ ,  $a$ , and  $\mathbf{R}$ . This can be problematic, especially when the dimension of the problem is large, and we know

that marginal likelihoods are quite sensitive to changes in the hyper-parameters; see O’Hagan & Forster (2004, Ch. 7). We therefore suggest an objective choice, based on the fractional matrix normal Wishart prior (20) applied to the Gaussian likelihood (12) with  $(n - n_0)$  observations and the same  $\hat{\mathbf{B}}$ ,  $\tilde{\mathbf{C}}$  and  $\tilde{\mathbf{R}}$  as the data. With this choice, the terms  $m(\mathbf{Y}_{\text{fa}_{\mathcal{D}}(j)})$  and  $m(\mathbf{Y}_{\text{pa}_{\mathcal{D}}(j)})$  in formula (27) can be computed from (22) provided that the condition  $|\text{fa}_{\mathcal{D}}(j)| = |\text{pa}_{\mathcal{D}}(j)| + 1 < n - p$  is satisfied. This condition guarantees a valid value for  $m(\mathbf{Y}_j | \mathbf{Y}_{\text{pa}_{\mathcal{D}}(j)}) = m(\mathbf{Y}_{\text{fa}_{\mathcal{D}}(j)}) / m(\mathbf{Y}_{\text{pa}_{\mathcal{D}}(j)})$  by granting a proper distribution to the marginal precision matrix  $\boldsymbol{\Omega}_{\text{fa}_{\mathcal{D}}(j) \text{fa}_{\mathcal{D}}(j) \cdot \overline{\text{fa}_{\mathcal{D}}(j)}}$ ; see section 4.2. In this way, formula (27) provides us with a valid marginal likelihood (product of  $q$  valid conditional marginal likelihoods given parent observations) for every DAG  $\mathcal{D}$  whose parent sets have size smaller than the number of observations minus the number of columns in the design matrix  $\mathbf{X}$  (number of predictors in the model plus one). The latter is a sparsity condition on the structure of the DAG, involving the maximal number of parents across vertices, which is quite reasonable in our intended application setting (eQTL analysis) as discussed in the Introduction.

## 5.2 Decomposable error structure

It is often appropriate to model the conditional independence structure of a set of variables in terms of an undirected graph; see Lauritzen (1996) for an authoritative exposition. This is for instance the approach followed in Cai *et al.* (2013) and Chen *et al.* (2013) for the analysis of genetical genomics data. With reference to the Gaussian multivariate regression model (11), this means that the precision matrix  $\boldsymbol{\Omega}$  of the response vector  $\mathbf{y}_i$  is constrained by an undirected graph  $\mathcal{G}$ : if an edge is missing between  $j$  and  $j'$  in  $\mathcal{G}$ , then  $\boldsymbol{\Omega}_{jj'} = 0$ . Equivalently,  $\mathbf{y}_i$  is Markov with respect to  $\mathcal{G}$ , that is, if  $j$  and  $j'$  are not joined by an edge in  $\mathcal{G}$ , the responses  $y_{ij}$  and  $y_{ij'}$  are conditionally independent, under the sampling distribution, given all remaining responses; in symbols  $y_{ij} \perp\!\!\!\perp y_{ij'} \mid \mathbf{y}_{i(\{1, \dots, q\} \setminus \{j, j'\})}$ ,  $\mathbf{B}, \boldsymbol{\Omega}$  (Drton & Perlman, 2004).

To enhance tractability, the undirected graph  $\mathcal{G}$  is often assumed to satisfy some conditions, such as *decomposability*; see for instance Bhadra & Mallick (2013). It is well known that a decomposable  $\mathcal{G}$  is Markov equivalent to some DAG (Andersson

*et al.*, 1997). Specifically, one can always well-number the vertices of  $\mathcal{G}$  and construct a directed version  $\mathcal{G}^<$ , which is a DAG Markov equivalent to  $\mathcal{G}$ ; see Lauritzen (1996, p. 18). It follows that the methodology developed in section 5.1 can also be applied to decomposable graphs, because the marginal likelihoods given by such methodology are invariant with respect to Markov equivalence. Indeed, the proof of Theorem 4 in Geiger & Heckerman (2002) directly carries over into our regression setting.

In practice, the marginal likelihood of the model defined by the decomposable graph  $\mathcal{G}$ ,  $m_{\mathcal{G}}(\mathbf{Y}) = m_{\mathcal{G}^<}(\mathbf{Y})$ , will be given by (27) with  $\mathcal{D} = \mathcal{G}^<$ . Since the parameter prior used to compute (27) satisfies global parameter independence,  $m_{\mathcal{G}^<}(\mathbf{Y})$  is readily seen to be  $\mathcal{G}^<$ -Markov; see for instance Cowell *et al.* (1999, sect. 9.4). Then  $m_{\mathcal{G}}(\mathbf{Y})$  is also  $\mathcal{G}$ -Markov, and thus it admits the representation

$$m_{\mathcal{G}}(\mathbf{Y}) = \frac{\prod_{C \in \mathcal{C}} m(\mathbf{Y}_C)}{\prod_{S \in \mathcal{S}} m(\mathbf{Y}_S)}, \quad (28)$$

where  $\mathcal{C}$  is the set of cliques, and  $\mathcal{S}$  the set of separators, of the decomposable graph  $\mathcal{G}$ ; see Lauritzen (1996). The explicit expression of each factor appearing in (28) can be deduced from (17) as explained in section 3.2.

In particular, when using the fractional matrix normal Wishart prior (20), one computes  $m(\mathbf{Y}_C)$  and  $m(\mathbf{Y}_S)$  in (28) by means of (22), with  $J = C$  and  $J = S$ , respectively, assuming  $|C| < n - p$  (hence  $|S| < n - p$ ) whenever  $C$  is a clique ( $S \subseteq C$  a separator) of  $\mathcal{G}$ . In this way, we cope with decomposable graphs whose clique sizes are smaller than the number of observations minus the number of predictors in the model. This is again a sparsity assumption on the graph, well-suited to our intended application setting, which grants a proper distribution to  $\boldsymbol{\Omega}_{CC \cdot \bar{C}}$  (hence to  $\boldsymbol{\Omega}_{SS \cdot \bar{S}}$ ); see section 4.2. We remark that formulae (28) and (22) generalize to the multivariate regression setup the results established by Carvalho & Scott (2009) for i.i.d. Gaussian observations with zero expectation. As a special case, formulae (28) and (22) also cover the i.i.d. Gaussian setup with unknown expectation.

## 6 Discussion

Motivated by covariate-adjusted covariance selection under sparsity, this paper proposes an objective Bayes method for computing the marginal likelihood of a multivariate regression model with normally distributed errors whose covariance matrix is constrained by a DAG. This represents an essential ingredient to obtain a posterior probability over the space of covariate-adjusted DAG models. Since the proposed method is invariant with respect to Markov equivalence, it can also be used to select covariate-adjusted decomposable models. Although we do not explicitly address variable selection, our results for the marginal likelihood can be used for Bayesian joint variable and covariance selection, as discussed in Bhadra & Mallick (2013).

In practice, as we remark at the beginning of section 3, variable selection is needed to apply our method whenever the total number of predictors  $p_*$  is comparable to, or larger than, the number of observations; this is a typical scenario in genetical genomics applications. Restricting our attention to models including only  $p \ll n$  predictors, so that our objective analysis becomes feasible, turns out to be adequate for settings where sparse models are of interest. For instance, the two simulations considered by Bhadra & Mallick (2013) have: i)  $p_* = 498$ ,  $q = 300$ , and  $n = 120$ , with  $p = 11$  for the actual data generating distribution; ii)  $p_* = 498$ ,  $q = 100$ , and  $n = 120$ , with  $p = 3$  for the actual data generating distribution. Similarly, their real data analysis (eQTL Analysis on Publicly Available Human Data) has  $p_* = 3125$ ,  $q = 100$ , and  $n = 60$ , with  $p = 1$  or  $p = 2$  identified as the most likely values.

Bhadra & Mallick (2013) currently derive their results for decomposable models under a weakly informative prior which requires to subjectively specify three scalar hyper-parameters. In particular, they use a hyper-inverse Wishart on  $\Sigma$  with scale parameter equal to the identity matrix multiplied by a constant. The latter proves to be crucial and need be fixed with care, because it acts as a global shrinkage parameter. Our objective prior, with its simple method for obtaining the marginal likelihood, should provide a useful alternative to their prior specification. On the other hand, our methodology for computing the marginal likelihood can also be implemented starting

from a single subjectively specified matrix normal Wishart prior under any complete DAG model, then applying the general results of section 3.2 in the context of DAG models as described in section 5.1. In this case, the sparsity conditions relating the sample size  $n$ , the number of predictors  $p$  and the maximal size of the cliques, which we had to impose to make our objective Bayes analysis possible, could be relaxed.

Finally, our method does not cope with non-decomposable undirected graphical models. Bayesian covariate-adjusted covariance selection in general undirected graphical models is beyond the scope of this paper, and will present the obvious challenge of providing an efficient method for computing the marginal likelihood; see Carvalho *et al.* (2007), Wang & Carvalho (2010), Lenkoski (2013). However, working within the class of decomposable graphs can still be very effective, even when the true graph is not decomposable; see Fitch *et al.* (2014) for asymptotic results on the posterior model probabilities, and for a high performing stochastic search of the model space.

## Acknowledgements

Work partially supported by a D1-grant from Università Cattolica del Sacro Cuore. The authors are grateful to Alberto Roverato for pointing out a useful reference.

## References

- Andersson, S. A., Madigan, D. & Perlman, M. D. (1997). On the Markov equivalence of chain graphs, undirected graphs, and acyclic digraphs. *Scand. J. Statist.* **24**, 81–102.
- Bhadra, A. & Mallick, B. K. (2013). Joint high-dimensional Bayesian variable and covariance selection with an application to eQTL analysis. *Biometrics* **69**, 447–457.
- Brem, R. B. & Kruglyak, L. (2005). The landscape of genetic complexity across 5,700 gene expression traits in yeast. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 1572–1577.



- Cai, T. T., Li, H., Liu, W. & Xie, J. (2013). Covariate-adjusted precision matrix estimation with an application in genetical genomics. *Biometrika* **100**, 139–156.
- Carvalho, C. & Scott, J. (2009). Objective Bayesian model selection in Gaussian graphical models. *Biometrika* **96**, 497–512.
- Carvalho, C. M., Massam, H. & West, M. (2007). Simulation of hyper-inverse Wishart distributions in graphical models. *Biometrika* **94**, 647–659.
- Chen, M., Ren, Z., Zhao, H. & Zhou, H. (2013). Asymptotically Normal and Efficient Estimation of Covariate-Adjusted Gaussian Graphical Model. *ArXiv e-prints* To appear in Journal of the American Statistical Association.
- Consonni, G. & La Rocca, L. (2012). Objective Bayes factors for gaussian directed acyclic graphical models. *Scandinavian Journal of Statistics* **39**, 743–756.
- Cowell, R. G., Dawid, P. A., Lauritzen, S. L. & Spiegelhalter, D. J. (1999). *Probabilistic networks and expert systems*. Springer, New York.
- Dawid, A. P. (1981). Some matrix-variate distribution theory: Notational considerations and a bayesian application. *Biometrika* **68**, 265–274.
- Dawid, A. P. (2003). Causal inference using influence diagrams: the problem of partial compliance. In P. Green, N. L. Hjort & S. Richardson, eds., *Highly structured stochastic systems*. Oxford Univ. Press, Oxford, pp. 45–81.
- Dawid, A. P. & Lauritzen, S. L. (1993). Hyper Markov laws in the statistical analysis of decomposable graphical models. *The Annals of Statistics* **21**, 1272–1317.
- DeGroot, M. H. (1970). *Optimal statistical decisions*. McGraw-Hill Book Co., New York.
- Drton, M. & Perlman, M. (2004). Model selection for Gaussian concentration graphs. *Biometrika* **91**, 591–602.

- Fitch, A. M., Jones, M. B. & Massam, H. (2014). The performance of covariance selection methods that consider decomposable models only. *Bayesian Anal.* **9**, 659–684.
- Geiger, D. & Heckerman, D. (2002). Parameter priors for directed acyclic graphical models and the characterization of several probability distributions. *Ann. Statist.* **30**, 1412–1440.
- Geisser, S. (1965). Bayesian estimation in multivariate analysis. *Ann. Math. Statist.* **36**, 150–159.
- Geisser, S. & Cornfield, J. (1963). Posterior distributions for multivariate normal parameters. *J. Roy. Statist. Soc. Ser. B* **25**, 368–376.
- Gupta, A. K. & Nagar, D. K. (2000). *Matrix variate distributions*. Chapman & Hall/CRC, Boca Raton, FL.
- Heckerman, D., Geiger, D. & Chickering, D. (1995). Learning bayesian networks: The combination of knowledge and statistical data. *Machine Learning* **20**, 197–243.
- Kuipers, J., Moffa, G. & Heckerman, D. (2014). Addendum on the scoring of Gaussian directed acyclic graphical models. *Ann. Statist.* **42**, 1689–1691.
- Lauritzen, S. L. (1996). *Graphical models*. Oxford University Press, Oxford.
- Lauritzen, S. L. (2001). Causal inference from graphical models. In *Complex stochastic systems (Eindhoven, 1999)*, vol. 87 of *Monogr. Statist. Appl. Probab.* Chapman & Hall/CRC, Boca Raton, FL, pp. 63–107.
- Lenkoski, A. (2013). A direct sampler for G-Wishart variates. *Stat* **2**, 119–128.
- Letac, G. & Massam, H. (2007). Wishart distributions for decomposable graphs. *Ann. Statist.* **35**, 1278–1323.
- Madigan, D., Andersson, S. A., Perlman, M. D. & Volinsky, C. T. (1996). Bayesian model averaging and model selection for Markov equivalence classes of acyclic digraphs. *Communications in Statistics - Theory and Methods* **25**, 2493–2519.

- Moreno, E. (1997). Bayes factors for intrinsic and fractional priors in nested models. Bayesian robustness. In Y. Dodge, ed., *L<sub>1</sub>-statistical procedures and related topics*. Institute of Mathematical Statistics, pp. 257–270.
- O’Hagan, A. (1995). Fractional Bayes factors for model comparison. *Journal of the Royal Statistical Society. Series B (Methodological)* **57**, 99–138.
- O’Hagan, A. & Forster, J. (2004). *Kendall’s advanced theory of statistics. Vol. 2B. Bayesian inference*. John Wiley & Sons, Ltd., Chichester.
- Pericchi, L. R. (2005). Model selection and hypothesis testing based on objective probabilities and Bayes factors. In D. Dey & C. R. Rao, eds., *Bayesian thinking: modeling and computation*, vol. 25 of *Handbook of Statistics*. Elsevier/North-Holland, Amsterdam, pp. 115–149.
- Press, S. J. (1982). *Applied multivariate analysis: Using Bayesian and frequentist methods of inference*. Krieger Publishing Company, Inc., Malabar, FL.
- Rothman, A. J., Levina, E. & Zhu, J. (2010). Sparse multivariate regression with covariance estimation. *J. Comput. Graph. Statist.* **19**, 947–962. Supplementary materials available online.
- Rowe, D. B. (2003). *Multivariate Bayesian statistics*. Chapman & Hall/CRC, Boca Raton, FL.
- Spiegelhalter, D. J., Dawid, A. P., Lauritzen, S. L. & Cowell, R. G. (1993). Bayesian analysis in expert systems. *Statist. Sci.* **8**, 219–283. With comments and a rejoinder by the authors.
- Sun, D. & Berger, J. O. (2007). Objective priors for the multivariate normal model. In J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. Smith & M. West, eds., *Bayesian Statistics 8 – Proceedings of the Eighth Valencia International Meeting*. Oxford University Press, pp. 525–554.

- Wang, H. & Carvalho, C. M. (2010). Simulation of hyper-inverse Wishart distributions for non-decomposable graphs. *Electron. J. Stat.* **4**, 1470–1475.
- Whittaker, J. (1990). *Graphical models in applied multivariate statistics*. Wiley, New York.
- Yin, J. & Li, H. (2011). A sparse conditional Gaussian graphical model for analysis of genetical genomics data. *Ann. Appl. Stat.* **5**, 2630–2650.